Online Learning and Integration of Complex Action and Word Lexicons for Language Grounding

Logan Niehaus and Stephen E. Levinson
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801 USA
Email: niehaus4@illinois.edu, sel@ifp.uiuc.edu

***Abstract*—This paper presents a computational framework for the development and integration of action and language capabilities through symbol grounding. Our approach is based around the fundamental technique of building lexicons of perceptual "simulators" which link noisy sensory experiences to internal symbolic representations. These sensory representations are paired with a basic model of association, allowing for the grounding of linguistic symbols directly in action knowledge – a grounding which is then exploited to bootstrap the development of more advanced capabilities. The performance of this computational framework is tested in the context of online tutoring scenarios using the iCub robotic platform. Such an experimental environment requires development of techniques and algorithms suited for incremental learning and real-time processing.**

I. Introduction

The field of cognitive robotics takes the position that cognition – and therefore language – is a necessarily embodied phenomenon. Under this view, language is not an isolated capability, but rather is part of an integrated cognitive faculty, which itself is shaped by biological, developmental and social factors. These principles are applied when exploring the fundamental problems concerning language, such as the problem of symbol grounding – how linguistic symbols get their meaning[1]. Cognitive robotics focuses on models which ground linguistic symbols in the sensorimotor experiences of an artificial agent, such as the iCub humanoid robot used in our experiments[2]. One particular aspect of the symbol grounding problem is the interaction between language and action. Current research continues to focus on the issues of representating linguistic and motor knowledge, linking these representations, and exploring how each representation informs and shapes the other[3]. Many approaches to these problems seek to develop computational techniques in line with basic principles of cognitive development: online learning, real-time processing, and sensory integration.

Previous work has already proven successful in developing computational frameworks for action-language integration that address some of these basic challenges. Many experiments have explored using artificial neural networks to ground linguistic labels for action words in the internal controllers used to produce those actions[4], [5]. Further experiments using these same types of models have demonstrated how this basic grounding mechanism can be exploited in order to bootstrap learning of more advanced compositional or hierarchical action-words through techniques like "grounding transfer"[6], [7]. In many of these architectures language grounding has been based upon action representations that are tightly integrated with simplistic language representations, and have not generally not dealt with issues of continuous exapansion or adaptation that are fundamental to online learning experiments.

Another popular approach is based around the use of statistical models to develop symbolic representations of action, while maintaining a separate model for associating action categories and words. Recent techniques have been developed for incrementally building action representations through observation of unmarked action data[8], [9], [10]. These models are able to adapt and expand to novel or unexpected inputs over the course of the experiment using online learning algorithms. Online methods have also been applied to learn associations between linguistic symbols and action categories[11]. While the action-specific nature of these representations do not easily lend themselves to grounding transfer applications, their underlying techniques could be abstracted to develop a framework for building representations of compositionally and hierarchically organized behaviors, as well as representations of language based in actual speech.

The goal of this work is to explore the application of statistical-model-based representations of action and language in the context of learning experiments similar to those presented in [6]. We achieve this through the development of a general representational model and learning techniques. Keeping in line with basic principles of cognitive development, algorithms should be capable of real-time, online learning and adaptation, and should be robust enough to handle the noisy sensory inputs of a real-world experimental environment. The rest of the paper is organized as follows. In Section II, the experimental scenario will be outlined and the proposed computational framework given. Section III presents the results of our experiments and discusses the their significance. Section IV concludes with comparisons to important related work, and makes note of areas of possible future improvement.

II. Computational Model

To begin, we first identify the goals of the model in terms of a simple 3-stage tutoring scenario, using [6] as a basic template. In the first stage, the tutor teaches the robot sets of basic actions and words. During the second stage, the robot

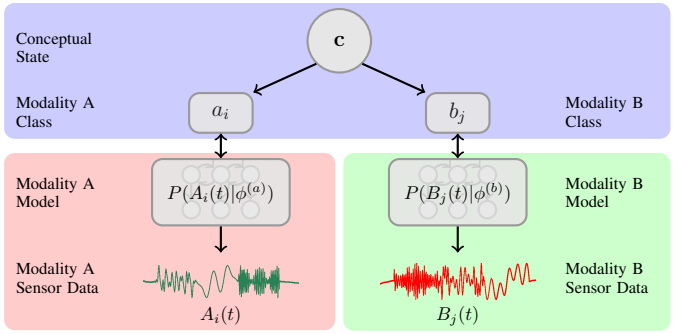


Fig. 1. Graphical representation of conceptual (blue) and perceptual (red/green) models and their relation to sensory observations in the proposed experiment.

makes action-word associations as the tutor provides speech labels for action demonstrations. The final stage consists of the tutor verbally describing a set of compositional actions in terms of their component actions. To be considered successful, our model should be able to learn the sets of actions and words, the associations between these two, and a set of complex actions and their names through verbal instruction only.

Drawing inspiration from the concepts of perceptual symbols and simulators[12], our overall approach to achieving these goals is the development of a two-level perceptual-conceptual model structure, depicted in Figure 1. Perceptual symbols indexing the various perceptual categories present in a given modality are linked to internal conceptual symbols within an associative memory. Perceptual simulators are generative models which represent perceptual categories and provide the link between the internal symbolic world and the noisy world of sensory experiences. To meet our experimental goals, we develop a set of techniques for constructing lexicons of perceptual simulators, based on the online sequence clustering methods explored in [8], [10].

A. Acquisition of Perceptual Lexicons

The first task is constructing a perceptual lexicon. Elements in the lexicon (perceptual simulators) may model items such as words (speech), actions (proprioception), visual objects/events, or any salient categories within a sensory stream. We have chosen to use the hidden Markov model (HMM) as the basis of our perceptual simulators, as it has seen wide application to both the modeling of speech and action. The algorithm for constructing the lexicon and learning the parameters of its individual models is based on the idea of treating a given sensory stream as a chain of perceptual events generated by a statistical model from that modality's lexicon.

Formally, we consider a lexicon to be a set of HMMs $\mathcal{K} = \{\phi_1, \ldots \phi_k, \ldots \phi_K\}$, each with order r_k, which learn the structure of a given sensory modality. We assume that a sensory stream $A(t)$ can be segmented into sub-sequences $A_i(t)$ that have been produced by a single model of the lexicon. In a real-world setting, $A_i(t)$ will be a finite length sequence $A_i(t) = \left[A_i^{(0)}, A_i^{(1)}, \ldots A_i^{(T)}\right]$. The elements of $A_i(t)$ may be either d-dimensional real-valued vectors, or discrete valued symbols, drawn from a dictionary of size d. Categorization involves finding the model which best fits the sequence:

$$a_i = \arg\max_{k \in \mathcal{K}} P(A_i(t)|\phi_k). \quad (1)$$

In a learning scenario such as ours, however, we do not know what these categories are, or even how many categories exist. The categories and their models must be learned as the robot experiences the world, and must be continually adjusted with each new experience.

Algorithm 1 Lexicon Creation Algorithm

$K \leftarrow 0$
while $A_i(t)$ **do**
  for $k = 1$ to K **do**
    $\mathcal{L}_k = (1/T_i) \log [P(A_i(t)|\phi_k)]$
    $\Lambda_k = \mathcal{F}(\mathcal{L}_k, \phi_k)$
  end for
  if $\{k \in K : \Lambda_k > \theta_0\} = \{\emptyset\}$ **then**
    Increment K by 1; Create new model ϕ_K
    $\phi_K = \text{train}(A_i(t), \phi_K)$
    $\mu(\phi_K) = \mathcal{L}_K$; $\sigma(\phi_K) = \sigma_0$
  else
    $\hat{k} = \arg\max_{\{k \in K : \Lambda_k > \theta_0\}} \mathcal{L}_k$
    $\phi_{\hat{k}} = \text{update}(A_i(t), \phi_{\hat{k}})$
    Update $\mu(\phi_{\hat{k}})$ and $\sigma(\phi_{\hat{k}})$ using $\mathcal{L}_k$
  end if
end while

To do this we propose the following general lexicon learning algorithm, outlined in Algorithm 1, which is based on a common competitive-learning principle. In words, the algorithm attempts to classify a new sequence $A_i(t)$ as one of current elements in the lexicon. If an existing model is determined to be close enough to the newly presented sequence, it is adjusted using the new data, while all other models remain unchanged. If no element satisfactorily "fits" the sequence, a new model is created and it is trained using the sequence. Training is the task of finding the parameters of a model that maximize the likelihood of the given data:

$$\phi^* = \arg\max_{\phi \in \Phi} P(A_i(t)|\phi). \quad (2)$$

While expectation-maximization[13] (EM) is the most popular HMM training method for (2), it requires all training samples to be processed simultaneously and is therefore not an ideal algorithm to use for online model adaptations. One alternative is to use the HMM's generative capabilities to balance old and new training data, while still applying EM[14]. Another popular class of approaches for online learning are stochastic gradient-based techniques[15].

However, before these methods can even be applied, we must first decide whether a given sequence belongs to an existing or previously unknown category. This detection of

novelty is made by applying a heuristic similar to the one presented in [16]. During classification, the log-likelihood value $\hat{\mathcal{L}}_k = \log[P(A_i(t)|\phi_k)]$ is calculated for each model. After calculating a length-normalized version of the log-likelihood $\mathcal{L}_k$, a transformation $\Lambda_k = \mathcal{F}(\mathcal{L}_k, \phi_k)$ is applied. This mapping is intended to account for variations in the goodness-of-fit achievable for an arbitrary class, and is given by:

$$\mathcal{F}(\mathcal{L}_k, \phi_k) = \int_{-\infty}^{\mathcal{L}_k} \mathcal{N}(x, \mu(\phi_k), \sigma^2(\phi_k))dx. \quad (3)$$

The parameters of the normal probability density function $\mathcal{N}(x, \mu(\phi_k), \sigma^2(\phi_k))$ are estimated for each lexical element based on the values of $\mathcal{L}_k$ calculated for previous training samples. After this mapping is applied, a threshold θ_0 is used to determine whether the lexical element "fits" the data.

B. Grounding of Conceptual Symbols via Perceptual Simulators

The result of (1) is the transformation of a set of noisy sensory sequences, $A_i(t)$, into a set of perceptual symbols, $a_i \in \{1, 2, \ldots K\}$. To learn the association between actions and words, we employ a latent variable model, where each perceptual symbol observation a_i is a sensory manifestation of an unobservable "concept". Each concept generates such symbols – potentially across many modalities – according to some probability mass function (PMF). The model parameters are the collection of these PMFs, expressed as a set of matrices, $\mathcal{O}$, with each matrix in the set representing a different modality:

$$[\mathcal{O}_a]_{m,k} = P(a = k|c = m). \quad (4)$$

The underlying conceptual state for ith observation, c_i, is an element of the set $\mathcal{C} = \{1, \ldots m, \ldots M\}$.

The goal of grounding language is to find the set of concepts, such that concepts which are likely to generate certain perceptual effects in a proprioceptive modality are also likely to generate the corresponding linguistic label in the speech modality. The model learns these associations without supervision or prior information, by observing pairs of cross-modal perceptual symbols, heuristically estimated to be conceptually linked. In our implementation, temporal cross-modal synchrony determines whether two symbols are linked. A measure of cross-modal synchrony between a pair of segments $A_i(t)$ and $B_j(t)$ is determined by dividing the duration of temporal overlap (referenced to a common clock) between the two segments by the duration of the shorter segment. If this synchrony measure exceeds a set threshold, the corresponding symbol pair $\{a_i, b_j\}$ is provided as a training sample to the associative memory model. The set of model parameters $\psi = \{\mathcal{O}_a, \mathcal{O}_b\}$ is estimated by maximizing the likelihood function for the set of N such:

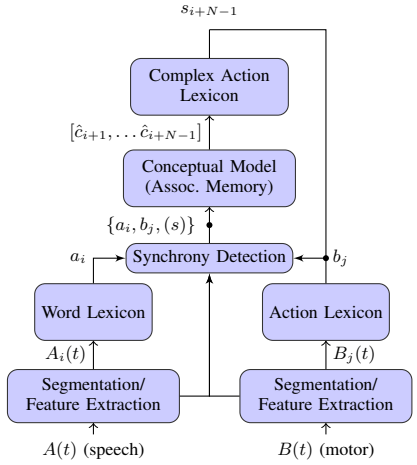


Fig. 2. Block diagram of the action-word learning framework, showing the flow of data between the various components.

$$\psi^* = \arg\max_{\psi \in \Psi} \prod_{n=1}^{N} \left(\sum_{c_n \in \mathcal{C}} P(a_{i_n}|c_n, \psi) P(b_{j_n}|c_n, \psi) \right). \quad (5)$$

Online algorithms for the parameter estimation problem in (5) have been successfully developed and applied to similar language-grounding experiments[17], [11]. After sufficient training, the hidden "concept" state can be estimated from an unpaired uni-modal input using:

$$\hat{c}_i \quad = \quad \arg\max_{m \in \mathcal{C}} [\mathcal{O}_a]_{m,a_i}. \quad (6)$$

C. Acquisition of Hierarchical Actions

After the initial phase of basic action-language learning, we would like the robot to be able to exploit this basic grounding to learn a compositional action. In practical terms, this would consist of a verbal instruction from the tutor of the form "wave [is] raise, left, right, left, right, lower", where "raise", "lower", "left", and "right" are all previously grounded basic action-words. However, before we can associate "wave" with this sequence, we must first produce a symbolic representation for it. Fortunately the same representational structure outlined above can be applied here as well, simply by adding another perceptual lexicon that takes estimated concept sequences from the associative memory model using (6). The output of the compositional lexicon can then be fed back into the associative memory as another perceptual variable, and used to ground the new label ("wave" in this case). A block diagram of the complete action-word learning architecture is shown in Figure 2.

III. Experiments in Action-Word Learning

As noted in the introduction, the target testbed for the presented framework is an online tutoring scenario between a human instructor and a real-world robot pupil. The goal of the experiments was to successfully complete the outlined

tasks in an environment which makes the similar demands of robustness to noise and real-time computational performance that are made on humans. The following experiments were performed using the iCub humanoid robot[2] and a standard desktop PC for computation.

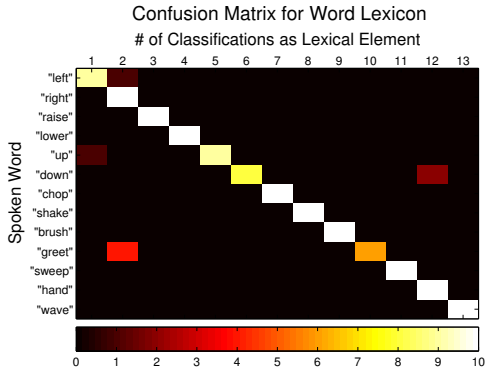
A. Experiment I: Incremental Acquisition of Perceptual Lexicons

The first goal was that of action and word lexicon acquisition. The lexicon learning algorithm presented above was evaluated on its ability to correctly categorize the various sensorimotor patterns presented to it – a task which includes determining not only category membership, but also the number of categories present. Here we tested the performance of the algorithm for modeling speech and motor modalities.

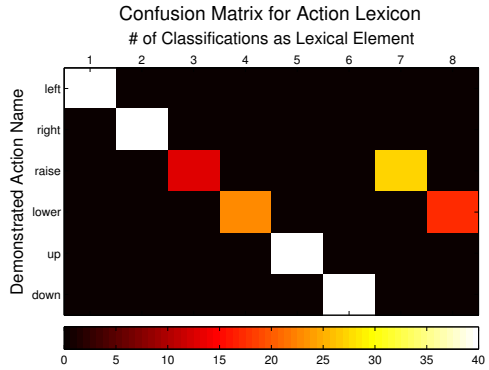
1) Word Lexicon Acquisition: Word acquisition performance was tested by presenting the lexicon learning algorithm with a semi-continuous stream of speech from a male speaker. The speech sample consisted of 13 English action words, each spoken 10 times in random order. A brief pause was taken between the pronunciation of each word for ease of segmentation. The audio stream was first transformed into a sequence of vectors of 13 Mel-frequency cepstral coefficients (MFCCs)[18], with each vector corresponding to roughly a $\tau_w = 25$ms window. The signal energy calculated over each window was low-pass filtered ($f_c = 1/(5\tau_w)$) and thresholded to separate speech segments. The sequence was then passed to a phonetic classifier consisting of a 10th-order HMM with multivariate Gaussian output distributions. The classifier was trained without supervision on a separate two-minute speech sample from the same speaker, using the k-means and EM algorithms.

Discrete-observation sequences corrsponding to internal state estimates of the phonetic classifier over each segment were produced using standard dynamic-programming methods, and presented to the speech lexicon. Initial training for newly created elements was done using the EM algorithm, with subsequent adaptations made using a stochastic gradient-descent algorithm[15]. Individual HMMs were of fixed size $r = 7$, with left-to-right topology. The novelty detection threshold θ_0 was set to 0.01 for this and all subsequent lexicon learning experiments. For the speech sample described above, the lexicon learner correctly estimated the size of the lexicon and produced a "ground-truth" confusion rate of 6.15%. Figure 3a presents a confusion matrix, which displays the number of times each utterance of a word was classified as a particular lexical element.

2) Action Lexicon Acquisition: The lexicon of primitive actions to be learned by the robot consisted of six basic motions: moving the hand left, right, up and down, as well as raising and lowering the hand to and from a position above its shoulder. These six actions were each presented 40 times in random order, via direct manual manipulation of the robot's right arm by the human tutor. Pauses were again taken between each demonstration to aid in segmentation. Cartesian position and velocity measurements (sampled at 50Hz) for



(a) Word confusion matrix



(b) Action confusion matrix

Fig. 3. Confusion matrices for word and action lexicon learning tasks. Ordinate labels are given only denote the "ground-truth" action/word categories intended by the tutor, and were not provided to the robot during training.

the end-effector were the sensory-level features presented to the algorithm. The L^2 magnitude of the end-effector velocity was low-pass filtered ($f_{cutoff} = f_{samp}/5$) and thresholded to segment the action sequences. Each sequence was preprocessed by subtracting the initial position from each vector in the sequence, resulting in actions that were expressed relative to current hand position.

Training of a new model for a novel sequence was performed by first initializing output means and covariances with the k-means algorithm, followed by application of the EM algorithm. Model adjustments were made using a modification of the generative technique presented in [14]. All models were of order $r = 4$. The resulting confusion matrix for the experiment is given in Figure 3b. While there was no confusion between lexical elements, the algorithm created 8 models to represent the 6 intended action categories demonstrated by the tutor. The multiple categories captured the different stylistic variations of a basic action. The consequences and significance of this particular result become clearer when integrated into the following language grounding scenarios.

B. Experiment II: Online Learning of Action-word Associations and Compositional Actions

In the next experiment, these uni-modal perceptual representations were integrated via an associative memory to produce a

simple, grounded linguistic faculty. We then sought to exploit this grounding by using the previously developed perceptual lexicon learning techniques in order to expand the range of learnable behaviors to compositionally and hierarchically organized actions. This two-stage experimental procedure consisted of an initial basic action-word tutoring phase, which then bootstrapped a second complex action-word learning phase.

1) Stage 1: Basic Action-word Tutoring: During the first stage the tutor manually demonstrated an action on the robot while simultaneously speaking the word describing the action. 100 such training samples for the 6 basic actions were randomly presented by the tutor in semi-continuous action and speech streams as before. The associative memory model parameters were incrementally updated using the same online training method of [17]. The number of columns for each modality's observation matrix within the associative memory model was dynamically expanded to match the current size of the corresponding lexicon, while the number of rows was fixed to $M = 6$ for this first stage.

2) Stage 2: Complex Action-word Learning via Initial Grounding: After the initial basic action-words demonstrations, the tutor was given the option of presenting the robot with verbal instructions in the form of fixed-script sentences describing a new, multi-step action sequence in terms of the basic component actions learned previously. These sentences consisted of the name for the new action followed by the sequence of component action names. The speech segment for this sentence was decoded using dynamic programming on an extended word lattice constructed from end-to-end concatenations of the individual lexicon HMMs[19]. The last $N-1$ symbols of the length-N decoded word-symbol sequence were converted to an estimated concept sequence $[\hat{c}_{i+1}, \ldots \hat{c}_{i+N-1}]$ using (6) and provided to a new, discrete-observation lexicon learner. The resulting lexicon classification symbol s was paired with the first word to produce the training sample $\{a_i, \emptyset, s\}$, with the symbol $\emptyset$ representing a "non-observation" placeholder symbol used for the inactive modality. Individual HMMs in this lexicon were created with a number of internal states equal to the length of the training sequence, up to a maximum of $r = 4$, and their parameters were trained using the EM algorithm.

In this second stage, both basic action-word demonstrations and sentence descriptions of the 5 multi-step action words given in Table I were presented. Sentences describing the first 4 complex actions in Table I in terms of the 6 initial basic actions, were presented in random order, 20 times each. After this, the final action was introduced and described 20 times. This fifth action, expressed in terms of two complex action-words itself, was chosen to demonstrate how new levels of scaffolding can be seamlessly constructed by the model. As soon as a sufficient number training samples for reliable estimation of the internal conceptual state have been given, any learned action-word grounding can be leveraged for multi-step action representation. However, this requires the development of a heuristic for handling the necessary dynamic exapansion of the concept model's state space as new words are added.

TABLE I
COMPLEX ACTION SEQUENCE DESCRIPTIONS

Name	Instruction Sequence
"wave"	"raise, left, right, left, right, lower"
"chop"	"raise, down, up, lower"
"shake"	"up, right, left, right, left, down"
"brush"	"right, left, right, left"
"greet"	"wave, shake"

In this experiment, our heuristic was simply to match the concept model order to that of the dynamically expanding word lexicon.

As in the initial stage, the associative memory model updated its parameters incrementally whenever a training sample was received. For the second stage, the internal state space was expanded to 11. The final estimates of the three observation matrices are shown in Figure 4. From this we can see how the latent variable structure captures action-word associations. Internal states having a high probability mass for a given action category were also likely to produce a word observation symbol corresponding to that action's linguistic label. Additionally, we see that even when the perceptual lexicons discovered a larger number of *perceptual* categories for a given modality than expected (as seen in the action lexicon learning experiment), they may still be mapped onto the same *conceptual* category when integrated with linguistic information, as was done here.

IV. DISCUSSION & CONCLUSION

The results of these experiments not only demonstrate the ability of the computational framework developed in this paper to satisfy the goals of the proposed action-word learning experiment, but also highlight the advantages afforded by the particular models and algorithms used. Issues of scalability and complexity often present in neural-network approaches[6], are mitigated through use of a separate concept-model that is robust to many of the structural aspects of a perceptual model that is capable of expansion and adaptation to novel or noisy inputs – as demonstrated in Experiment II. By generalizing a basic online sequence clustering approach developed for actions[8], [10], we produced a framework capable of incrementally learning a word representation based in actual speech – an aspect of sensory grounding for linguistic symbols not explored in [6], [7].

We were also able to extend the results of previous experiments employing statistical models of language grounding to translate sentences to sequences of primitive actions[11], by applying our generalized lexicon learning algorithm to build a representation of complex actions from these sentences. This enabled us to transfer meaning from basic action-word groundings to new linguistic symbols describing compositionally or hierarchically organized actions. Additionally, the use of estimated *concept* symbols as the basis for this higher-order representation allowed the model to preserve knowledge of hierarchical organization of complex actions. Architectures like those in [7] have typically grounded complex actions

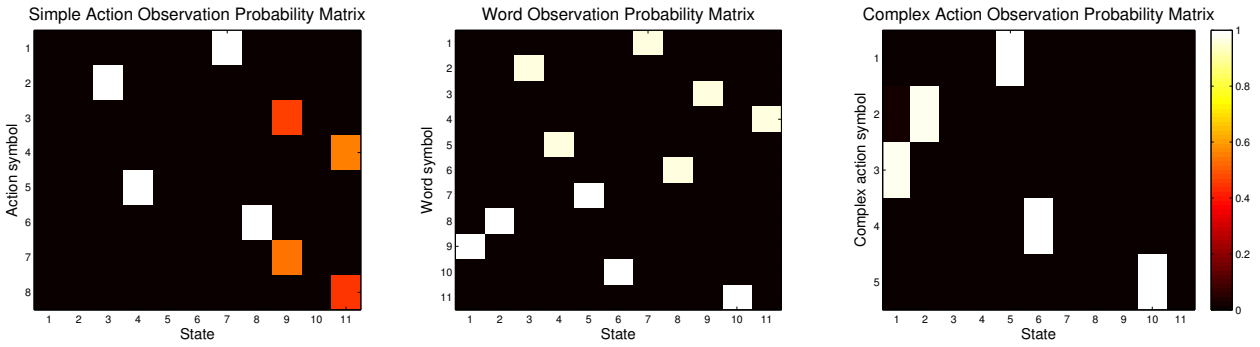


Fig. 4. Observation symbol matrices for the action, word, and complex action lexicons. Each column is the PMF for a single state of the internal model. Observation probabilities corresponding to the "no-observation" symbol for each modality are not shown.

directly in the primitive action set, and thus are unable to capture such organizational structure.

Many of the constraining assumptions and heuristics required by this framework are specific examples of more general open problems faced by similar kinds of experiments. We believe that future work seeking to address these challenges will serve to further demonstrate the advantages of applying our framework to model the sensorimotor grounding of language. Linguistic representations rooted in actual speech allow us to exploit sensory integration using basic Bayesian techniques to correct word comprehension errors, in a way that resembles hypothesized *action-compatibility effects*[20]. The application of a uniform set of models and algorithms for building perceptual lexicons affords us the possibility to adapt proven automatic segmentation techniques[8], allowing for the relaxation of constraints placed on experimental scenarios.

In conclusion, this paper has presented a simple computational framework for the learning and grounding of action-words in terms of sensorimotor experience, and demonstrated its capabilities in the context of an online human-robot tutoring scenario. We showed that the model was not only able to incrementally build speech and motor representations and integrate them to produce a basic grounded linguistic faculty, but also could exploit this multi-modal interaction to construct more complex representations of compositionally and hierarchically organized behaviors. In doing so, we extended the range of capabilities for statistical model-based approaches to action-language integration, while addressing many of the challenges faced by alternate approaches.

REFERENCES

[1] S. Harnad, "The symbol grounding problem," *Physica D*, vol. 42, no. 1-3, pp. 335 – 346, 1990.

[2] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The iCub humanoid robot: an open platform for research in embodied cognition," in *PerMIS: Performance Metrics for Intelligent Systems Workshop.*, Washington DC, USA, August 2008, pp. 50–56.

[3] A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, F. Nori, L. Fadiga, B. Wrede, K. Rohlfing, E. Tuci, K. Dautenhahn, J. Saunders, and A. Zeschel, "Integration of action and language knowledge: A roadmap for developmental robotics," *Autonomous Mental Development, IEEE Transactions on*, vol. 2, no. 3, pp. 167 –195, Sept. 2010.

[4] Y. Sugita and J. Tani, "Learning semantic combinatoriality from the interaction between linguistic and behavioral processes," *Adaptive Behavior*, vol. 13, no. 1, pp. 33–52, 2005.

[5] V. Tikhanoff, A. Cangelosi, and G. Metta, "Integration of speech and action in humanoid robots: iCub simulation experiments," *Autonomous Mental Development, IEEE Transactions on*, vol. 3, no. 1, pp. 17 –29, March 2011.

[6] A. Cangelosi and T. Riga, "An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots," *Cognitive Science*, vol. 30, no. 4, pp. 673–689, 2006.

[7] F. Stramandinoli, D. Marocco, and A. Cangelosi, "The grounding of higher order concepts in action and language: A cognitive robotics model," *Neural Networks*, vol. 32, no. 0, pp. 165 – 173, 2012.

[8] D. Kulic, W. Takano, and Y. Nakamura, "Online segmentation and clustering from continuous observation of whole body motions," *Robotics, IEEE Transactions on*, vol. 25, no. 5, pp. 1158 –1166, oct. 2009.

[9] T. Inamura, I. Toshima, H. Tanie, and Y. Nakamura, "Embodied symbol emergence based on mimesis theory," *The International Journal of Robotics Research*, vol. 23, no. 4-5, pp. 363–377, 2004.

[10] W. Takano and Y. Nakamura, "Humanoid robot's autonomous acquisition of proto-symbols through motion segmentation," in *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*, dec. 2006, pp. 425 –431.

[11] ——, "Statistically integrated semiotics that enables mutual inference between linguistic and behavioral symbols for humanoid robots," in *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, may 2009, pp. 646 –652.

[12] L. W. Barsalou, "Perceptual symbol systems," *Behavioral and Brain Sciences*, vol. 22, no. 04, pp. 577–660, 1999.

[13] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Soc. Series B*, vol. 1, no. 39, pp. 1–38, 1977.

[14] S. Calinon and A. Billard, "Incremental learning of gestures by imitation in a humanoid robot," in *Proc. of the ACM/IEEE Intl. Conf. on Human-Robot Interaction*, 2007, pp. 255–262.

[15] V. Krishnamurthy and G. Yin, "Recursive algorithms for estimation of hidden Markov models and autoregressive models with Markov regime," *IEEE Trans. on Information Theory*, vol. 48, pp. 458–476, 2002.

[16] H. Brandl, B. Wrede, F. Joublin, and C. Goerick, "A self-referential childlike model to acquire phones, syllables and words from acoustic speech," in *Development and Learning, 2008. ICDL 2008. 7th IEEE International Conference on*, Aug. 2008, pp. 31 –36.

[17] K. Squire and S. Levinson, "HMM-based semantic learning for a mobile robot," *IEEE Trans. on Evolutionary Computation*, vol. 11, pp. 199–212, 2007.

[18] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357 – 366, 1980.

[19] S. Levinson, *Mathematical Models for Speech Technology*. New York, NY: John Wiley and Sons Ltd., 2005.

[20] A. Glenberg and M. Kaschak, "Grounding language in action," *Psychonomic Bulletin Review*, vol. 9, pp. 558–565, 2002.